

# Validation of Chain-of-Thought prompt structuring to improve a language model's capability to extract new clinical concepts from clinical text messages

Matthew Farmer, PhD, RN<sup>1</sup> , Kimberly Powell, PhD, RN, FAMIA<sup>1</sup> , & Mihail Popescu, PhD<sup>2</sup> 

*Collaborators: 2 reviewers*

<sup>1</sup> Sinclair School of Nursing, University of Missouri, Columbia, MO, USA

<sup>2</sup> Department of Biomedical Informatics, Biostatistics, and Medical Epidemiology, School of Medicine, University of Missouri, Columbia, MO, USA

Accepted by 2 of 2 reviewers

## Conflicts of Interest

The authors declare no conflicts of interest.

## Publishing History

Submitted May 31 2024

Accepted January 20 2026

Published March 13 2026

## Corresponding Author

Matthew Farmer  
msfppy@missouri.edu



Open Access



Peer-Reviewed



Creative Commons

Artificial Intelligence Clinical Applications

Computer & Software Science Healthcare

Large Language Models

## Abstract

The rapid improvement in generative language model capabilities and adoption has increased interest in their application for healthcare, particularly in extracting meaningful information and language from unstructured text. This study investigates the validity of Chain-of-Thought (CoT) prompting on a small language model's ability to extract clinical concepts from nursing home staff text messages. We focus on concepts within the Age-Friendly Health Systems 4M framework (What Matters, Medication,



Mentation, and Mobility). We evaluated the classification performance of four locally-run language models on an expert-annotated dataset of 860 text messages, using classification metrics to compare with or without CoT prompting. Our results demonstrate that CoT prompting improves the capabilities of Gemma 2 9b in extracting expressions related to the 4M framework. However, there was not an improvement observed in the Medication concept. CoT continues to demonstrate potential in improving language model capabilities in the clinical domain, particularly from unstructured text data. This has implications for ontology development, information retrieval, and clinical decision support. Additional contexts or data sources would improve the generalizability of these findings.

**Keywords:** age-friendly health systems, computer science, gerontology, natural language processing

## Introduction

Prior to the public release of ChatGPT in November 2022, Wei et al. (2022) published their foundational paper demonstrating that the inclusion of intermediate reasoning steps, or chain-of-thought (CoT), significantly improved the capabilities of generative language models in the areas of arithmetic, commonsense, and symbolic reasoning tasks (Wei et al., 2022). This growing field of research is, in part, due to the rapid expansion and improved capabilities of generative language models in multiple domains, prompting individuals to examine the scope and depth of large language model capabilities (AL-Smadi et al., 2022, Loh, 2023, Banerjee et al., 2024, Chakraborty et al., 2024, Sahoo et al., 2024).

The purpose of the following study is to provide validation of CoT prompting on the capabilities of a small language model tasked to extract expressions of clinical concepts from unstructured text messages sent to and by clinicians surrounding transfer events in nursing homes. These clinical concepts are found in the evidence-based Age-Friendly Health



Systems framework including the 4M's; "What Matters", "Medication", "Mentation" and "Mobility" (Figure 1: Mate et al., 2021). As part of a larger study (Powell et al., 2023), the authors are analyzing text messages sent by nursing home staff to begin development of an ontology supporting this 4M framework. The difficulty of extracting clinical concepts from unstructured text messages is due to the non-standard usage of terms, vocabulary, and contextual placement of conversational text. Our team tested multiple natural language processing techniques to overcome this barrier, including the use of generative language models. Our aim for this report is to describe the prompting techniques we utilized to achieve improved performance in extracting clinical concepts from our text message dataset compared to our expert-annotated "Gold Standard" dataset.

## Methods and Materials

### Model selection

A small selection of language models capable of running on a local device were chosen to extract the 4M concepts and compared to an expert annotated "Gold Standard" dataset. Models were compared using classification metrics including accuracy, precision, recall, F1, receiver-operator characteristics area-under-curve (ROC-AUC), and Cohen's Kappa ( $\kappa$ ).

The model with highest "off-the-shelf" capabilities was chosen for additional testing and fine tuning. The models tested included: (1) Gemma 2 9b q4\_0; (2) Llama 3.1 8b q4\_0; (3) Llama 3.1 8b q8\_0; (4) Mistral Nemo 12.2b q4\_0; and (5) Gemma 2 2b fp16. Model parameters were standardized across models for consistency using the Langchain python package. For thought generation, temperature was set to 0.5 with 500 max tokens. For extraction, temperature was 0.2, repetition penalty = 1.2, repetition context size = 20, top p = 0.95, top\_k = 40, and 250 max tokens. Structured, parsed response parameters were set to temperature = 0.0, repetition penalty = 1.2, repetition context size = 20, top p = 0.95, and 50 max tokens. Initial model evaluation tested the performance of a simple extraction task for each model using a system prompt shown in Figure 1



and a prompt for each of the 4Ms, asking the model to extract the correct terms. After initial model testing, the highest performing model was chosen for testing of additional prompt engineering including CoT, model parameter changes, and fine-tuning. Model extractions were compared to the annotated dataset with classification model metrics including: Precision, Recall, F1-Score, Area-under-Curve (AUC), and Cohen's Kappa. Bootstrap confidence intervals were calculated on model metric comparisons with 1000 bootstrap interactions per test and  $\alpha = 0.05$ .

### **Data source and standard**

The data source includes 21,357 text messages sent to and by healthcare workers from 12 nursing homes participating in the Missouri Quality Initiative, a CMS-funded innovation and care coordination project that began in 2012 (Powell et al., 2023). The "Gold Standard" dataset consisted of a sample of 860 text messages extracted from the full 21,209 messages and annotated by two clinical experts familiar with the 4M framework. Sampling was accomplished by randomly choosing 40 transfer events and annotating any messages sent up to two weeks prior to the event for 4M content. Annotations were provided independently, then the annotators aligned any discrepancies through discussion of any differing extractions. Additional details of the sampling and annotation methods have been previously reported (Powell et al., 2023). This sample was utilized to assess the model capabilities.

### **Chain-of-thought prompting**

The initial extraction task included a simple prompt instructing the models to make a list of words or phrases consistent with individual 4M concepts. Context (see Figure 1) was provided to the model as a system prompt briefly describing the conceptual definitions of each 4M.

In our chain-of-thought prompting test, we added an additional task prior to term extraction informing the model to provide an explanation of why any words or phrases belong to the individual 4M concept (Xie, 2024). No changes were made to the other prompts or contextual messages.



\*The four components of the 4M Framework are:

**M1 – What Matters:** Understanding and acting on the specific health goals and care preferences of the older adult.

- Examples of M1 Include:
  - Discussions about code status, DNR orders, and advanced directives.
  - Words or phrases related to palliative care, hospice, or end-of-life care.
  - Inclusion of family and caregivers in decision-making and care planning.
  - Discussions about quality of life, values, and preferences.
  - Aligning treatment and care plans with the older adult’s goals and preferences.

**M2 – Medication:** Ensuring that medications do not interfere with the other aspects of the 4M Framework and are optimized for the older adult’s health.

- Examples of M2 Include:
  - Medication reconciliation and review.
  - Deprescribing or reducing the number of medications.
  - Addressing polypharmacy and potential drug interactions.
  - Ensuring appropriate medication use and adherence.
  - Monitoring for side effects and adverse drug reactions.

**M3 – Mentation:** Addressing mental health, including cognitive function, depression, and delirium.

- Examples of M3 Include:
  - Screening for cognitive impairment and dementia.
  - Assessing and managing depression and anxiety.
  - Preventing, identifying, and managing delirium.
  - Addressing sleep disorders and other mental health issues.
  - Promoting brain health and cognitive function.

**M4 – Mobility:** Promoting physical activity and preventing falls to maintain or improve function.

- Examples of M4 Include:
  - Assessing and addressing gait and balance issues.
  - Preventing falls and fall-related injuries.
  - Encouraging physical activity and exercise.
  - Promoting mobility and independence.
  - Addressing barriers to physical activity and mobility.

Given this background, please extract and categorize concepts related to the 4M Framework from the following unstructured text messages. For each message, identify and label the relevant concepts under the appropriate “M” category in JSON format. Do not include names or any personal information in the reference that can identify the individuals in the message, instead use resident/family or exclude any personal information.

Some messages are not applicable to the 4M framework, in which case you should return an empty string for each concept.

These include greetings, salutations, expressions of gratitude, emojis, out-of-office messages, adding or removing people from the etc.

There are also messages that are very short (sometimes just a number or a letter) and do not contain any relevant information. In these cases, you should return an empty string for each concept.”

**Figure 1.** System Prompt - 4M context



## Results

Gemma 2 9b (9 billion parameters) (Farabet and Warkentin, 2024), Gemma 2 2b, Llama 3.1 8b (Dubey et al., 2024), and Mistral Nemo 12.2b (Mistral AI, 2024) were chosen for initial testing. Models were utilized on a local device for testing to eliminate the risk of sending personal health information online. Model parameters shared between the models included temperature (randomness) = 0.0, top\_p = 0.9, top\_k = 40, and seed (random state) = 418. Gemma 2 9b was the highest performing model for the initial extraction test compared to other models. Chain-of-thought (CoT) prompting was then added. An example of the output is shown in Figure 2. Classification results, displayed in Table 1, improved significantly in 6 metrics in the What Matters and Mobility domains. No significant improvements were achieved in the Mentation domain.

In the Medication domain, the non-chain-of-thought (NonCoT) prompting approach outperformed the chain-of-thought (CoT) model in key classification metrics, with significant differences observed in precision (NonCoT: 0.704 vs. CoT: 0.650; difference: -0.054, 95% CI: -0.096 to -0.015,  $p=0.006$ ), F1 score (NonCoT: 0.816 vs. CoT: 0.781; difference: -0.035, 95% CI: -0.066 to -0.005,  $p=0.020$ ), and Cohen's Kappa (NonCoT: 0.772 vs. CoT: 0.726; difference: -0.046, 95% CI: -0.086 to -0.007,  $p=0.016$ ), indicating agreement and balanced performance without CoT's added reasoning steps. Recall showed no significant difference (CoT: 0.979 vs. NonCoT: 0.972;  $p=0.758$ ), and AUC was comparable (CoT: 0.937 vs. NonCoT: 0.945;  $p=0.266$ ), suggesting that for medication-related extractions from clinical text, the simpler NonCoT model maintains high discriminative power without the overhead of CoT.

Overall, both models correctly extracted 3053 of 3440 (88.8%) classifications, when the models disagreed on classification, CoT was correct 44.9% of the time. The strongest significant differences by effect size were found in: What Matters - Recall (CoT better by 0.33), F1 (CoT better by 0.14), AUC (CoT better by 0.12), Cohen's Kappa (CoT better by 0.12); and Mobility - Recall (CoT better by 0.13).



**Table 1.** Model classification metrics with 95% CI bootstrap significance.

Domain	Metric	CoT Score	Non-CoT Score	Difference	Lower Bootstrap CI	Upper Bootstrap CI	Bootstrap $p$	Significant	Better Model
What Matters	Precision	0.529	0.616	-0.0869	-0.1785	0.0045	0.0640	No	No Difference
What Matters	Recall	0.702	0.376	0.3262	0.2340	0.4113	0.0000	Yes	CoT
What Matters	F1	0.604	0.467	0.1367	0.0569	0.2207	0.0000	Yes	CoT
What Matters	AUC	0.790	0.665	0.1249	0.0777	0.1689	0.0000	Yes	CoT
What Matters	Cohen Kappa	0.513	0.391	0.1212	0.0279	0.2129	0.0120	Yes	CoT
Medication	Precision	0.650	0.704	-0.0537	-0.0960	-0.0146	0.0060	Yes	Non-CoT
Medication	Recall	0.979	0.972	0.0069	-0.0163	0.0331	0.7580	No	No Difference
Medication	F1	0.781	0.816	-0.0352	-0.0658	-0.0049	0.0200	Yes	Non-CoT
Medication	AUC	0.937	0.945	-0.0084	-0.0223	0.0065	0.2660	No	No Difference
Medication	Cohen Kappa	0.726	0.772	-0.0460	-0.0859	-0.0073	0.0160	Yes	Non-CoT
Mentation	Precision	0.613	0.611	0.0018	-0.1024	0.1027	0.9520	No	No Difference
Mentation	Recall	0.792	0.688	0.1042	-0.0004	0.2308	0.1280	No	No Difference
Mentation	F1	0.691	0.647	0.0439	-0.0466	0.1417	0.3860	No	No Difference
Mentation	AUC	0.881	0.831	0.0502	-0.0076	0.1130	0.1100	No	No Difference
Mentation	Cohen Kappa	0.670	0.625	0.0453	-0.0501	0.1492	0.3920	No	No Difference
Mobility	Precision	0.643	0.714	-0.0714	-0.2251	0.0755	0.3600	No	No Difference
Mobility	Recall	0.491	0.364	0.1273	0.0417	0.2222	0.0000	Yes	CoT
Mobility	F1	0.557	0.482	0.0748	-0.0127	0.1725	0.0920	No	No Difference
Mobility	AUC	0.736	0.677	0.0593	0.0175	0.1074	0.0000	Yes	CoT
Mobility	Cohen Kappa	0.531	0.459	0.0721	-0.0203	0.1739	0.1320	No	No Difference



## Discussion

Our study demonstrates that chain-of-thought prompting enhances the ability of a small language model, Gemma 2 9b, to extract expressions and 4M clinical concepts from unstructured text messages in the areas of What Matters and Mobility. This finding aligns with previous research by Wei et al., which empirically demonstrated these improvements in their 2022 study (Wei et al., 2022). While CoT improved the performance of extraction in many metrics, the non-CoT model performed significantly better in the medication domain in precision, F1, and Cohen's Kappa. This suggests that adding a reasoning step may introduce complexity within the Medication concept. In many cases, CoT overanalyzed an extraction where there were no elements to extract. To help illustrate areas where the CoT model misclassified messages, we present 3 cases.

CASE M04251: Non-CoT Correct, CoT Wrong

FULL MESSAGE:

"FYI Upon assessment this morning resident was noted to have periods of increased fatigue and weakness with some unresponsiveness. Irregular heartbeats. Vital signs are as follows: 110/86, 98.1, p89-52, r21, O2@ 91% on RA Assisted to bed and no s/sx of pain or distress noted at this time."

PREDICTIONS:

- Gold Standard: 0 (no medication)
- CoT Model: 1 ✗
- Non-CoT Model: 0 ✓

EXTRACTED REFERENCES:

- Gold Standard: '[]'
- CoT Extract: '['vital signs']'

Model Reasoning: "The text mentions vital signs, which can be indicators of medication effectiveness or side effects."

- Non-CoT Extract: '[]'

CASE M04272: Non-CoT Correct, CoT Wrong

FULL MESSAGE:

"TSH results"

PREDICTIONS:

- Gold Standard: 0 (no medication)



- CoT Model: 1 ✘
- Non-CoT Model: 0 ✔

EXTRACTED REFERENCES:

- Gold Standard: '[]'
- CoT Extract: '['TSH']'

Model Reasoning: "TSH is a blood test that can indicate thyroid function, which can be relevant to medication management."

- Non-CoT Extract: '[]'

CASE M00461: Non-CoT Correct, CoT Wrong

MESSAGE (truncated):

"Wanted to inform you that [...] urology office provided a Covid-19 PCR test that was showed that [...] was positive for Covid-19. We understand that this was a rapid test. We have isolated..."

PREDICTIONS:

- Gold Standard: 0 (no medication)
- CoT Model: 1 ✘
- Non-CoT Model: 0 ✔

EXTRACTED REFERENCES:

- Gold Standard: '[]'
- CoT Extract: '['rapid']'

Model Reasoning: "The text mentions a rapid COVID-19 test, which could be related to medication management as some medications may interact with the virus or affect testing results."

- Non-CoT Extract: '[]'



Message: "FYI Upon assessment this morning resident was noted to have periods of increased fatigue and weakness with some unresponsiveness. Irregular heartbeats. Vital signs are as follows: 110/86, 98.1, p89-52, r21, 02@ 91% on RA Assisted to bed and no s/sx of pain or distress noted at this time..."

Response:

M1\_reasoning: The text mentions the resident's periods of increased fatigue and weakness which could indicate a change in their well-being or goals. Additionally, the phrase "Please advise sir" suggests a need for communication and potentially understanding the resident's wishes.

M1\_ref: ["increased fatigue", "weakness", "Please advise sir"]

M2\_reasoning: "The text mentions "irregular heartbeats" which could indicate a need for medication management.

M2\_ref: ["irregular heartbeats"],

M2\_reasoning: The resident is described as having "periods of unresponsiveness", suggesting fluctuations in their mental state and awareness.

M3\_ref: ["unresponsiveness", "periods"],

M4\_reasoning: The text states the resident was "assisted to bed" indicating a need for help with mobility. ✨

M4\_ref: ["assisted to bed"]

**Figure 2.** Extraction example with chain-of-thought (CoT).

The Gold Standard dataset, limited to annotated text messages, limits the generalizability of our results. Further, our study is focused on a specific set of clinical concepts in the application of text extraction from unstructured text messages. The effect of CoT prompting may vary when applied to other clinical domains or data sources.



## **Implications for ontology development, information retrieval, and clinical decision support**

Our examination of CoT and our ongoing language processing research with the Age-Friendly Health Systems 4M framework (Powell et al., 2023) have provided us with insight into the limitations of language models for ontology development and information retrieval. Despite the improvements of model performance to extract words and phrases from unstructured text using CoT, language models fail to adhere to ontological commitments or align with consensus among subject-matter experts (Neuhaus, 2023). This underscores a key limitation of current generative models in healthcare informatics: their sensitivity to prompt structure for nuanced, context-dependent concepts, rather than rote extraction. However, the relative gains from CoT (e.g., 15-20% improvement in F1 scores for Mobility and What Matters categories, as detailed in Results) are not trivial. The results provide evidence for hybrid approaches in resource-limited environments, such as fine-tuning smaller models combined with structured prompting to approximate expert consensus without extensive retraining. Additional research into the utilization of language models in concept extraction and ontology engineering is ongoing. For clinical decision support systems relying on accurate extractions and categorization of unstructured text in age-friendly health systems, it appears that medication concepts are trained into small language models (without CoT) but more complex concepts such as What Matters and Mobility indicate the need for CoT reasoning.

## **Funding Information**

Research reported in this publication was supported by the National Institute on Aging of the National Institutes of Health under Award Number R01AG078281. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.



## Author Contributions

Matthew Farmer: Conceptualization, software, resources, formal analysis, investigation, methodology, visualization, writing - original draft

Kimberly Powell: Data curation, resources, supervision, funding, writing - review, project administration

Mihail Popescu: Supervision, writing - review

## Data Availability

Due to the personal health information contained in this study, data are not available to share.

## Supplemental Information

This study was completed in Python (version 3.12.4) using the following libraries: (1) Pandas (pandas, 2024); (2) json; (3) LangChain (Chase, 2022); and (4) typing. Pydantic (Colvin et al., 2024) classes were utilized to generate structured outputs for use in analysis and additional study.

## Transparent Peer Review

Results from the Transparent Peer Review [can be found here](#).

## Recommended Citation

Farmer, M., K. Powell, and M. Popescu. 2026. Validation of Chain-of-Thought prompt structuring to improve a language model's capability to extract new clinical concepts from clinical text messages. *Stacks Journal*: 26001. <https://doi.org/10.60102/stacks-26001>.



## References

- AL-Smadi, M. 2023. ChatGPT and beyond: The generative AI revolution in education. arXiv preprint arXiv:2311.15198. <https://doi.org/10.48550/arXiv.2311.15198>.
- Banerjee, S., P. Dunn, S. Conard, and A. Ali. 2024. Mental health applications of generative AI and large language modeling in the United States. *International Journal of Environmental Research and Public Health* 21: 910. <https://doi.org/10.3390/ijerph21070910>.
- Chakraborty, C., S. Pal, M. Bhattacharya, and A. Islam. 2024. ChatGPT or LLMs can provide treatment suggestions for critical patients with antibiotic-resistant infections: a next-generation revolution for medical science?. *International Journal of Surgery* 110: 1829-1831. <http://dx.doi.org/10.1097/JS9.0000000000000987>.
- Chase, H. 2022. LangChain (Version 0.1.22). <https://github.com/langchain-ai/langchain>.
- Colvin, S., E. Jolibois, H. Ramezani, A. Garcia Badaracco, T. Dorsey, D. Montague, S. Matveenko, M., Trylesinski, S. Runkle, D. Hewitt, A. Hall, and V. Plot. 2024. Pydantic Validation (Version 2.8.2). <https://docs.pydantic.dev/latest/>.
- Dubey, A., A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, and A. Goyal. 2024. The llama 3 herd of models. arXiv e-prints, pp.arXiv-2407. <https://doi.org/10.48550/arXiv.2407.21783>.
- Farabet, C. and T. Warkentin. 2024. Google launches Gemma 2, its next generation of open models. Accessed September 27, 2024. <https://blog.google/technology/developers/google-gemma-2/>.
- Loh, E. 2023. ChatGPT and generative AI chatbots: challenges and opportunities for science, medicine and medical leaders. *BMJ Leader* 8: e000797. <https://doi.org/10.1136/leader-2023-000797>.
- Mate, K., T. Fulmer, L. Pelton, A. Berman, A. Bonner, W. Huang, and J. Zhang. 2021. Evidence for the 4Ms: interactions and outcomes across the care continuum. *Journal of Aging and Health* 33: 469-481. <https://doi.org/10.1177/0898264321991658>.
- Mistral AI. 2024. Mistral NeMo. Accessed September 27, 2024. <https://mistral.ai/news/mistral-nemo/>.
- Neuhaus, F. 2023. Ontologies in the era of large language models—a perspective. *Applied Ontology* 18: 399-407. <https://doi.org/10.3233/AO-230072>.
- pandas. 2024. Python Data Analysis Library (Version 3.12). Accessed January 22, 2024. <https://pandas.pydata.org/>.
- Powell, K. R., M. Popescu, S. Lee, D. R. Mehr, and G. L. Alexander. 2023. Examining the use of text messages among multidisciplinary care teams to reduce avoidable hospitalization of nursing home residents with dementia: protocol for a secondary analysis. *JMIR Research Protocols* 12: e50231. <https://doi.org/10.2196/50231>.
- Sahoo, P., A. K. Singh, S. Saha, V. Jain, S. Mondal, and A. Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. arXiv preprint arXiv:2402.07927. <https://doi.org/10.48550/arXiv.2402.07927>.
- Wei, J., X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. V Le, and D. Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *In Advances in Neural Information Processing Systems*, edited by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, vol. 35. Curran Associates, Inc. [https://proceedings.neurips.cc/paper\\_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf).
- Xie, Z. 2024. Order matters in hallucination: Reasoning order as benchmark and reflexive prompting for large-language-models. arXiv preprint arXiv:2408.05093. <https://doi.org/10.48550/arXiv.2408.05093>.